MedJus: A Cloud-based LLM Application to Expedite Decision Making for Health Judicialization in Brazil

Bruno Padilha[1], **Jacson Venâncio de Barros**[2], **Giovanni Guido Cerri**[3], **João Eduardo Ferreira**[1]

[1]Instituto de Matemática e Estatística - Universidade de São Paulo (IME-USP)
São Paulo – SP – Brazil

[2]Amazon Web Services (AWS)
Brazil

[3]Faculdade de Medicina da Universidade de São Paulo (FMUSP)
São Paulo – SP – Brazil

brunopadilha@usp.br, jvbarros@amazon.com, giovanni.cerri@hc.fm.usp.br, jef@ime.usp.br

Abstract. *The increasing number of processes regarding health judicialization in Brazil has been contributing to a major overloading in the judiciary system. In spite of inherent difficulties in adapting Large Language Models (LLMs) to niche yet critical domains (e.g hallucinations, user preference alignment, response grounding), deploying these models to assist specialist users such as judges and medical doctors introduces several infrastructural challenges. In this preliminary work, we explore how the cloud native services provided by Amazon Web Services (AWS) can be employed to overcome these challenges in MedJus: a secure, scalable, and serverless cloud-native LLM application to assist and support decision making in health judicialization diligence for Brazilian cases.*

1. Intoduction

The rapid evolution of cloud computing has fundamentally transformed the deployment and scalability of artificial intelligence (AI) applications, particularly in the realm of Large Language Models (LLMs). As modern applications increasingly rely on LLMs to perform complex, domain-specific reasoning, the underlying computational infrastructure has become as critical as the algorithmic architectures themselves. Deploying state-of-the-art LLMs requires vast amounts of dynamic computational power, low-latency processing capabilities, and highly scalable storage—demands that are uniquely satisfied by modern cloud environments.

This paper highlights the indispensable role of cloud computing in facilitating the development of LLM-based applications for health judicialization in Brazil. We explore how cloud native solutions, such as managed services, dynamic resource provisioning, distributed computing, and specialized hardware accelerator clusters (GPUs/TPUs) facilitates state-of-the art model deployment while mitigating latency constraints to enable real-time workflows. Furthermore, we also leverage cloud infrastructure to ensure the data privacy required for handling sensitive medical data and legal proceedings. By abstracting the complexities of infrastructure, cloud platforms empower developers to build secure, scalable, and highly available AI applications that can be seamlessly integrated

with services and platforms currently in use by government departments and offices. To this end, we present our progress so far on developing MedJus: a cloud-native platform built on Amazon Web Services (AWS) to assist and support decision making in health judicialization diligence.

The remainder of this paper is structured as follows: Section 2 reviews the state of LLM deployment in cloud environments; Section 3 presents an overview of MedJus; Section 4 presents the cloud-based architecture supported by Amazon Web Services; and Section 5 presents our concluding remarks and directions for future research.

2. Related Work

To better contextualize the infrastructural demands of health judicialization applications, first we give an overview of current state of Large Language Models (LLMs) in health-legal domains and subsequently discuss the role of cloud ecosystems in academic research and system development in this area.

2.1. LLM usage in health-legal domain

The integration of Natural Language Processing (NLP) and, more recently, LLMs into the legal and medical domains has been extensively documented in recent literature. In the legal sector, models are frequently deployed for contract analysis, precedent retrieval, and legal document summarization. In the medical field, LLMs are utilized for medical images interpretation [Wang et al. 2025], clinical evaluation and decision support [Shool et al. 2025] [Jeong et al. 2025], patient data variability [Joshi et al. 2026], navigating heterogeneous data source [Campos et al. 2025] and etc. Nonetheless, the intersection of these two fields presents additional computational and architectural challenges. Applications in this niche must seamlessly integrate clinical evidence (e.g. the efficacy of a high-cost drug) and legal frameworks (e.g. local legislation and jurisprudence). While academic researchers have successfully demonstrated the theoretical viability of fine-tuning open-source LLMs, external data contextualization with RAG, or utilizing prompt engineering to align responses, in these cross-domain tasks only the most sophisticate LLMs to date are capable of meeting their intricate reasoning requirements [Kant et al. 2025, Berger et al. 2025, Zhou et al. 2025, Li et al. 2024].Thus, the transition from local, isolated software to interactive high availability system infrastructure is quickly becoming mandatory [Trajanoski and Karadimce 2025] once consumer grade hardware is not yet powerful enough to run such models (Figure 1).

2.2. Cloud Computing Ecosystems for Academic Research

One of the main selling points of cloud providers is offering LLM access and management as a service (MaaS). They abstract much of the complexities of provisioning GPU clusters, managing API rate limits, and configuring load balancers. This allows researchers and software developers to focus on application logic and domain-specific features, such as appropriate data segmentation for Retrieval-Augmented Generation (RAG) and agent intelligence to make better use of contextual information. This work is being built on top of AWS due to maturity of software development services and the model agnostic capabilities of AWS Bedrock for LLM inference.

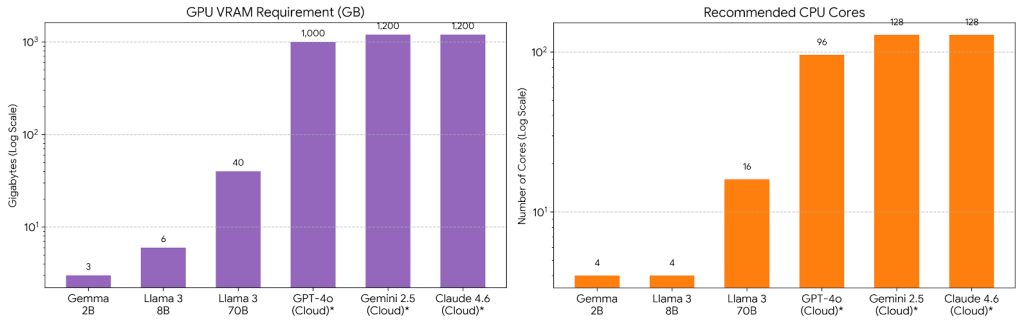

Figura 1. Comparison of system requirements (GPU VRAM and CPU Cores) between quantized open-weight models suitable for local inference (e.g., Gemma 2B, Llama 3) and the estimated server-side infrastructure required to run state-of-the-art models (e.g., GPT-4o, Gemini 2.5, Claude 4.6).

3. MedJus: Answers focused on health judicialization in Brazil

The essential challenge in health judicialization is grounding the LLM's output to verified trusted sources and localized data. Furthermore, recurring updates to the Unified Health System (SUS) clinical protocols or the latest jurisprudence from the Brazilian Supreme Federal Court (STF) makes relying solely on a model's pre-trained weights unfeasible. To overcome this limitations, the most straightforward approach is to enhance the LLM generation capabilities with contextual information for this specific domain at inference time. This can be accomplished with Retrieval-Augmented Generation (RAG) and clever prompt engineering. In MedJus, an specialized AI agent is responsible to orchestrate contextual retrieval from multiple RAG databases and dynamic prompt generation based on user input. Figure 2 depicts an user querying about use cases of a high-cost medication named *Pembrolizumabe*. Responses are always grounded on data from sources as enatjus [CNJ 2024], conitec [MS 2026], pubmed [NLM (US) 2026] and jurisprudence from state and federal courts. MedJus is currently in test phase collecting feedback from specialized users.

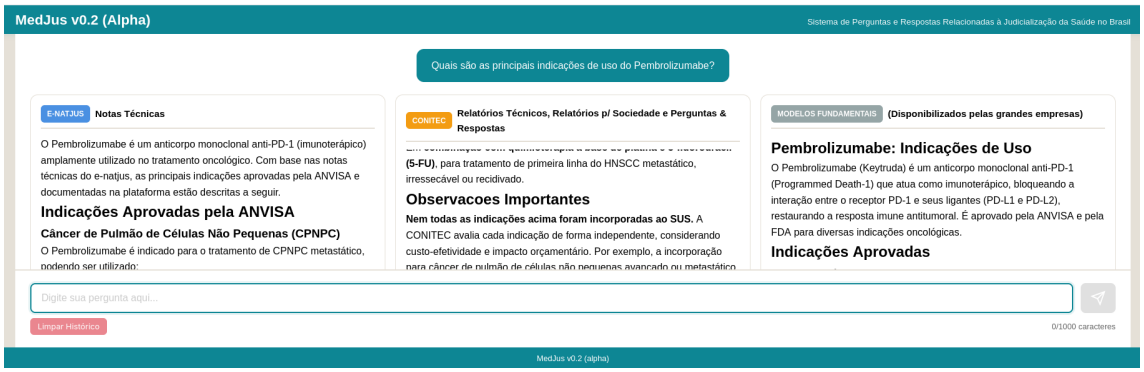

Figura 2. MedJus screenshot

4. Cloud Architecture

In Figure 3 we present the architectural details of MedJus using AWS. Collected data comes in various formats, for example, pdf files, html and json. This data is kept in a S3 bucket to facilitate data ingestion by SageMaker AI, a fully managed AWS service to

write machine learning code and interact with foundation models provided by Bedrock. Our data processing pipeline includes extracting structured information, either directly (e.g. html or json) or converting pdf file to text with tools such as Docling, segment this data into smaller chunks guided by specialist knowledge, computing embeddings with Bedrock and storing then in S3vectors for RAG. The core application leverage the Bedrock API to fulfill Agent requests to foundation models, coordinate tool calling and to generate the final output. On the development side, we adhered to Continuous Integration and Continuous Delivery/Deployment (CI/CD) practice [Chen 2024], a modern approach to delivery code changes in a more reliable and predictable manner in serverless cloud environments. It contemplates automated build routines upon commit with fallback to last working version when build fails. In AWS, we use CodePipeline, CodeCommit, CodeBuild, ECR, ECS and Fargate services to manage ci/cd pipeline. Other important services that handles network traffic, region availability, content delivery, usage monitoring, user authentication and etc. were omitted to simplify this presentation.

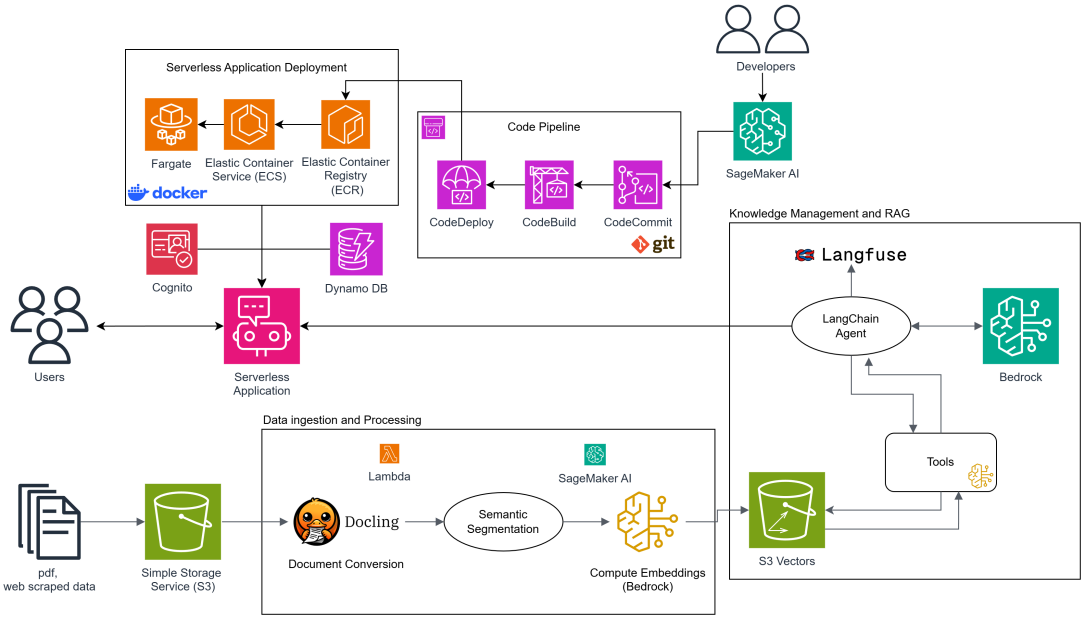

Figura 3. MedJus architecture in AWS

5. Conclusions and Future Work

Through the proposed architecture, we highlighted how managed services such as Amazon Bedrock abstract the complexities of LLM orchestration, while Amazon SageMaker AI and S3 Vectors provide a highly scalable, serverless pipeline for processing heterogeneous data and executing Retrieval-Augmented Generation (RAG). MedJus is a tool being developed to support legal decisions in health-legal domains. It is currently in test phase with a limited number of specialist users including judges and medical doctors to collect feedback to further optimize response content, preference alignment and to validate our context semantic segmentation approach, which will be explored in detail in a forthcoming publication.

Acknowledgment

This work received support from FAPESP (project 2021/11905-0: Centro de Ciência, Tecnologia e Desenvolvimento para Inovação em Medicina e Saúde), AWS Cloud Credit for Research program and DNX Brasil (dnxbrasil.com).

Artificial Intelligence Usage Declaration

We have used Artificial Intelligence (Google Gemmini 2.5 Pro) to generate Figure 1 and to obtain the data presented in this figure.

Referências

Berger, A., Khanna, S., Sparrenberg, L., Deußer, T., Berghaus, D., and Sifa, R. (2025). Reasoning llms in the medical domain: A literature survey. In *2025 IEEE 12th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. IEEE.

Campos, E. M., Conrado, R. C., Traina Jr, C., Traina, A. J., and Cazzolato, M. T. (2025). Assessing large language models for structuring patient records. In *Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 1–7. SBC.

Chen, T. (2024). Challenges and opportunities in integrating llms into continuous integration/continuous deployment (ci/cd) pipelines. In *2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*, pages 364–367. IEEE.

CNJ (2024). Sistema nacional de pareceres e notas técnicas (e-natjus). Acessado em: 12 mar. 2026.

Jeong, J., Kim, S., Pan, L., Hwang, D., Kim, D., Choi, J., Kwon, Y., Yi, P., Jeong, J., and Yoo, S.-J. (2025). Reducing the workload of medical diagnosis through artificial intelligence: A narrative review. *Medicine*, 104(6):e41470.

Joshi, S., Mehta, M., Maniar, S., Wang, M., and Singh, V. K. (2026). Performance of large language models under input variability in health care applications: Dataset development and experimental evaluation. *JMIR AI*, 5:e83640.

Kant, M., Nabi, S., Kant, M., Scharrer, R., Ma, M., and Nabi, M. (2025). Towards robust legal reasoning: Harnessing logical llms in law. *arXiv preprint arXiv:2502.17638*.

Li, S. S., Balachandran, V., Feng, S., Ilgen, J. S., Pierson, E., Koh, P. W., and Tsvetkov, Y. (2024). Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888.

MS (2026). Portal da comissão nacional de incorporação de tecnologias no sistema Único de saúde (conitec). Acessado em: 12 mar. 2026.

NLM (US) (2026). Pubmed. Acessado em: 12 mar. 2026.

Shool, S., Adimi, S., Saboori Amleshi, R., Bitaraf, E., Golpira, R., and Tara, M. (2025). A systematic review of large language model (llm) evaluations in clinical medicine. *BMC Medical Informatics and Decision Making*, 25(1):117.

Trajanoski, S. and Karadimce, A. (2025). Comparative analysis of large language models: On-premise architectures vs. cloud-based deployments. *Preface to Volume 5 Issue 2 of the Journal of University of Information Science and Technology "St. Paul the Apostle"–Ohrid*, 5(2):48.

Wang, P., Lu, W., Lu, C., Zhou, R., Li, M., and Qin, L. (2025). Large language model for medical images: A survey of taxonomy, systematic review, and future trends. *Big Data Mining and Analytics*, 8(2):496.

Zhou, S., Xie, W., Li, J., Zhan, Z., Song, M., Yang, H., Espinoza, C., Welton, L., Mai, X., Jin, Y., et al. (2025). Automating expert-level medical reasoning evaluation of large language models. *npj Digital Medicine*.